# Assignment Based Subjective Questions

1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: We may deduce the following regarding the categorical factors impact on the dependent variable ('cnt'), which indicates the demand for shared bikes, based on our examination of the dataset:

- Season :

The season has a major impact on the demand for bikes. We can see from the box plots that demand for bikes tends to be stronger during some seasons like summer and fall than during other like winter and spring. This could be because riding is a more alluring alternative during certain seasons due to the more favourable weather.

- Year:

Between 2018 and 2019 there was a rise in the demand for bikes. The box plot illustrate this, showing that the median number of bikes hired in 2019 is more than in 2018. This might be explained by the system's increasing uptake over time.

- Holiday:

When compared to non-holidays, the demand for motorcycles is somewhat lower on holidays. The box plot demonstrates how vacations tend to have lower median and overall distributions for bike rentals. One reason for this could be because fewer individuals go to work or participate in usual activities on vacations.

- Working Day:

On working days as opposed to non-working days, there is a greater demand for motorcycles. The box plot displays a working day distribution with a greater median count. This may be because more people are using bikes to commute throughout the workday.

- Weather Situation:

The demand for bikes is significantly influenced by the weather. The most popular weather circumstances for bikes are clear or partly overcast (category 1), while the least popular weather conditions for bikes are severe rain or snow (categories 3 and 4). Given that fewer people ride bikes in bad weather, this makes obvious sense.

Bike demand was dispersed across several categories and showed a central tendency when box plots were used to compare category factors with the target variable, cnt. Important findings included:

- Increased demand during the summer and fall.
- Demand that was higher in 2019 than in 2018.
- Reduced demand throughout the holidays.
- Increased demand during the working day.
- Most demand when the weather is clear.

2) Why is it important to use drop_first = True during dummy variable creation?

Ans: It's crucial to use {drop_first=True} while creating dummy variables in order to prevent multicollinearity and guarantee that the final dataset is appropriate for regression analysis. This is why it's so important to do this:

- Avoiding the Dummy Variable Trap:

You would obtain ( k ) dummy variables while creating dummy variables for a category feature with ( k ) levels. Nevertheless, multicollinearity —a situation in which the dummy variables have a high degree of correlation with one another—occurs when all ( k ) dummy variables are included in a regression model. We refer to this problem as the "dummy variable trap."

- How 'drop_first=Ture' Helps:

You can remove one dummy variable from each set of category variables by setting {drop_first=True}. This method eliminates the redundancy, which helps against falling into the dummy variable trap. The category that was deleted is now used as the benchmark for comparing the other categories.

By using {drop_first=True}, you can make sure that redundant dummy variables don't induce multicollinearity in your regression model. As a result, regression coefficients are more consistent and easy to understand, as each coefficient shows the impact of the associated category relative to the reference category.

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: We may look at the pair-plot and correlation matrix to see which numerical variable has the highest correlation with the target variable (cnt). The pair-plot will provide us a visual depiction of these associations, while the correlation matrix will provide us with a numerical measure of correlation.

The correlation coefficients between cnt and other numerical variables are visible to you via the correlation matrix. The variable most strongly correlated with bike demand is the one with the highest absolute correlation coefficient (positive or negative) with cnt.

Because individuals are more inclined to hire bikes in pleasant weather, the variable temp (temperature) often has a strong association with bike rental counts (cnt). The precise outcome, though, will vary depending on the dataset you choose.

```
             temp     atemp       hum  windspeed       cnt
temp     1.000000  0.987672  0.126963  -0.157944  0.627494
atemp    0.987672  1.000000  0.139988  -0.183643  0.631066
hum      0.126963  0.139988  1.000000  -0.248489 -0.100659
windspeed -0.157944 -0.183643 -0.248489  1.000000 -0.234545
cnt      0.627494  0.631066 -0.100659  -0.234545  1.000000
```

The measure that feels like temperature, or atemp, is most closely correlated with bike demand, as seen by its greatest correlation of 0.631066 with cnt.

4) How did you validate the assumptions of linear regression after building the model on the training set?

Ans: To guarantee the validity and dependability of the model, it is essential to validate the assumptions underlying linear regression. The main tenets of linear regression and the techniques employed to support them are as follows:

1. Linearity: There should be a linear connection between the predictors and the target variable.
2. Independence: Observations ought to stand alone from one another.
3. Homoscedasticity: At every level of the predictor variables, the residuals (errors) should have a constant variance.
4. Normality: A normal distribution should be seen in the residuals.
5. No Multicollinearity: There shouldn't be a strong correlation between the predictors

Steps To Validate Assumptions:

- Linearity:

To determine whether the predictors and the target variable have a
linear relationship, look at their scatter plots.
Plot the expected values against the observed values to check if a
45-degree line forms.

- Independence:

This relates more to the experiment's design or the procedure used to get the data. The Durbin-Watson test may be used to determine if residuals show autocorrelation for time series data.

- Homoscedasticity:

Plot the residuals against the values that were fitted, or anticipated. It is homoscedastic if the residuals are distributed randomly down the horizontal axis.
A heteroscedasticity pattern (such as a funnel shape) is indicated.

- Normality:

Plot the residuals' histogram and check to see if it resembles a normal distribution.
To compare the residual distribution to a normal distribution, use a Q-Q plot, also known as a quantile-quantile plot.
Run a statistical test, such as the Shapiro-Wilk test.

- No Multicollinearity:

For every predictor, determine the Variance Inflation Factor (VIF). High multicollinearity is indicated by VIF values greater than 5 or 10.

5)Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: We must examine the coefficients of the final linear regression model and their statistical significance in order to identify the top three features that substantially contribute to understanding the demand for shared bikes. The characteristics that have the most effects on the target variable (cnt) are those with the greatest absolute values of coefficients, provided that they remain statistically significant.

The coefficients and p-values for each feature may be seen in the model.summary() output. Here's what we're seeking for:

- High absolute value of coefficients: Indicates a stronger impact on the dependent variable.
- Low p-values(typically < 0.05): Indicates statistical significance.

Identifying the Top 3 Features

According to the summary, the top three characteristics are as follows: they have significant p-values ($p < 0.05$) and the highest absolute coefficients:

Temperature (temp): p-value < 0.0001, coefficient = 4531.5789
Coefficient = 2092.6281, p-value < 0.0001 for the year 2019 (yr_2019).

Season Fall (season_fall): p-value < 0.0001, coefficient = 2415.2730

The factors that have the greatest absolute influence on the target variable are as follows:

Fall season is associated with a higher demand for bikes compared to other seasons, probably because of favorable weather conditions.
Temperature has a strong positive effect on bike demand; as temperatures rise, more bikes are rented.
Year 2019 shows a significant increase in bike rentals compared to 2018, reflecting overall growth in the bike-sharing system.

# General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans:
Linear Regression Algorithm

A linear equation is used in the statistical technique of linear regression to represent the connection between a dependent variable (target) and one or more independent variables (predictors). It is among the most straightforward and widely applied predictive modeling strategies.

Objectives:

- Predicting Values: Estimate the value of the dependent variable based on the values of the independent variables.
- Understanding Relationships: Determine the strength and nature of the relationship between the dependent and independent variables.

Key Concepts:

- Dependent Variable (Y): The variable you are trying to predict.

- Independent Variables (X1, X2, ..., Xn): The variables you use to make predictions.

- Linear Equation: The relationship is modeled using a linear equation of the form:
  Y = beta_0 + beta_1X_1 + beta_2X_2 + ldots + beta_nX_n + epsilon

  - ( Y ) is the dependent variable.
  - ( X_1, X_2, ldots, X_n ) are the independent variables.
  - ( beta_0 ) is the intercept.
  - ( beta_1, beta_2, ldots, beta_n ) are the coefficients (weights) for the independent variables.
  - ( epsilon ) is the error term.

## Steps in Linear Regression:

- Data Collection: Gather data containing the dependent and independent variables.
- Exploratory Data Analysis (EDA): Understand the data using summary statistics and visualizations.
- Data Preprocessing: Handle missing values, encode categorical variables, and scale features if necessary.
- Model Building: Fit the linear regression model to the training data.
- Model Evaluation: Assess the model's performance using metrics such as R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
- Residual Analysis: Check the residuals to validate the assumptions of linear regression.
- Prediction: Use the model to make predictions on new data.

## Assumptions of Linear Regression

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: The residuals are independent

3. Homoscedasticity: The residuals have constant variance at every level of the independent variables.
4. Normality of Residuals: The residuals are normally distributed.

By adhering to the assumptions and appropriately testing the model, linear regression is a potent and interpretable method for modeling the connection between a dependent variable and one or more independent variables. It may yield insightful information and precise predictions.

2) Explain the Anscombe's quartet in detail.

Ans:
Anscombe's Quartet:

Four datasets comprise Anscombe's Quartet; they share virtually identical simple descriptive statistics, but their distributions and visual representations on a graph are significantly different. Francis Anscombe, a statistician, created it in 1973 to highlight the value of graphical data analysis.

Purpose:

Anscombe's Quartet aims to demonstrate that visualizing data is essential for comprehending its actual nature and that depending only on summary statistics (such as mean, variance, and correlation) can be deceptive.

The Four Datasets:

Here are the summary statistics that are identical across the four datasets:

Mean of x: 9.0
Mean of y: 7.5
Variance of x: 11.0
Variance of y: 4.125
Correlation between x and y: 0.816
Linear regression line: $y = 3.0 + 0.5x$

Despite these identical statistics, the datasets differ significantly when plotted.

Importance of Anscombe's Quartet:

- Graphical Analysis: Emphasizes the importance of visualizing data before drawing conclusions based on statistical measures.
- Misleading Statistics: Demonstrates that identical statistical properties can lead to different data distributions.
- Model Diagnostics: Underlines the need for proper model diagnostics and residual analysis in regression modeling.

Anscombe's Quartet serves as a potent reminder that understanding the complete narrative underlying data requires more than just statistical features. In data analysis, visualization is essential because it may highlight patterns that summary statistics might miss.

3) What is Pearson's R?

Ans: The linear correlation between two variables is expressed as Pearson's R, or Pearson correlation coefficient. It measures how well a straight line can capture the relationship between two variables. This measure has a range of -1 to 1, and it is represented by the symbol (r).

Formula

The formula for Pearson's correlation coefficient ( r ) is:

r = frac{sum (x_i - bar{x})(y_i - bar{y})}{sqrt{sum (x_i - bar{x})^2 sum (y_i - bar{y})^2}} ]

Where:

- ( x_i ) and ( y_i ) are the individual sample points.
- ( bar{x} ) and ( bar{y} ) are the means of the ( x ) and ( y ) variables, respectively.

- ( sum ) denotes the summation over all data points.

Properties:

- Symmetry: ( r(x, y) = r(y, x) ).
- Range: ( −1 leq r leq 1 ).
- Unitless: The correlation coefficient is a dimensionless measure, meaning it does not depend on the units of the variables.
- Linearity: It measures the strength and direction of a linear relationship between two variables.

Application:

1. Statistical Analysis: Commonly used in hypothesis testing to determine the strength of the relationship between two continuous variables.
2. Economics: To measure the correlation between variables such as inflation and unemployment.
3. Finance: To assess the relationship between the returns of different stocks or assets.
4. Medical Research: To examine the association between health outcomes and exposure factors.

Limitations:

1. Only Linear Relationships: Pearson's R only measures the strength and direction of a linear relationship. It does not capture non-linear relationships.
2. Sensitive to Outliers: Outliers can significantly affect the correlation coefficient.
3. Assumes Normality: Pearson's R assumes that the variables are normally distributed, which may not always be the case in real-world data.

In conclusion, Pearson's R is an effective method for measuring linear correlations between two variables, but it must be utilized with caution due to its assumptions and limitations.

4) What is Scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Ans: Scaling is a data preprocessing step used to adjust the range of features in your dataset so that they can be compared on a common scale. It is particularly important for machine learning algorithms that compute distances or assume data to be on a similar scale.

Why Scaling is Performed:

1. Improves Convergence : For algorithms like gradient descent, scaling can speed up convergence.
2. Enhances Model Performance: Many machine learning algorithms, such as k-nearest neighbors (KNN) and support vector machines (SVM), are sensitive to the scale of the data.
3. Prevents Dominance: Features with larger ranges can dominate those with smaller ranges, skewing the results.
4. Equal Weightage: Ensures that all features contribute equally to the analysis.

Types od Scaling

1. Normalisation(Min-Max Scaling)
2. Standardisation(Z-Score Scaling)

- Normalised Scaling:

Normalisation scales the data to a fixed range, usually [0,1] or [-1,1].

- Standardised Scaling:

Standardisation scales the data to have a mean of 0 and a standard deviation of 1.

Normalization (Min-Max Scaling):
Scales data to a fixed range [0, 1] or [-1, 1].
Sensitive to outliers.

Ensures all features are on the same scale.

Standardization (Z-score Scaling):
Centers data around mean 0 and standard deviation 1.
Less sensitive to outliers.
Assumes Gaussian distribution of the data.

When to Use Which

Normalization: Use when you know the data has a bounded range or when you are using algorithms that require bounded inputs.
Standardization: Use when the data follows a normal distribution or when the machine learning algorithm assumes that the data is normally distributed.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: One metric for determining if multicollinearity exists among a group of multiple regression variables is the Variance Inflation Factor (VIF). When independent variables exhibit significant correlation, a phenomenon known as multicollinearity arises, which makes it challenging to estimate the links between predictors and the dependent variable.

Why VIF Value can be infinite

One predictor variable is a precise linear combination of one or more of the other predictor variables when the VIF value is infinite, indicating perfect multicollinearity. As a result, the regression model's variable becomes unnecessary.

Causes of Infinite VIF:

1. Exact Linear Relationships: If any predictor variable is an exact linear combination of one or more other predictor variables, VIF will be infinite. For example, if ( X_3 = 2X_1 + 3X_2 ), then ( X_3 ) is perfectly collinear with ( X_1 ) and ( X_2 ).

2. Duplicate Variables: Including duplicate or identical variables in the regression model.
3. Dummy Variable Trap: Including all dummy variables for a categorical variable without dropping one category to avoid redundancy.

Dealing with Infinite VIF:

1. Remove Redundant Variables: Identify and remove perfectly collinear variables.
2. Check for Dummy Variable Trap: Ensure that for categorical variables, one category is excluded to avoid multicollinearity.
3. Principal Component Analysis (PCA): Transform the correlated variables into a set of uncorrelated components.

6)What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A graphical tool called a Q-Q plot (Quantile-Quantile plot) can be used to determine if a dataset adheres to a given theoretical distribution, most frequently the normal distribution. Plotting the dataset's quantiles versus the theoretical distribution's quantiles is what it does. The data is substantially regularly distributed if the points on the Q-Q plot roughly follow a straight line.

Construction of a Q-Q Plot

Sort Data: Arrange the sample data in ascending order.

Determine Theoretical Quantiles: Calculate the quantiles for the theoretical distribution (e.g., normal distribution).

Plot Data: Plot the sample quantiles on the x-axis and the theoretical quantiles on the y-axis.

Evaluate Linearity: Assess the linearity of the plotted points. If they lie along a straight line, the data distribution matches the theoretical distribution.

Use and Importance in Linear Regression

In the context of linear regression, the Q-Q plot is primarily used to assess the normality of the residuals, which is one of the key assumptions in linear regression.

Steps to Use a Q-Q Plot in Linear Regression

Fit the Model: Fit the linear regression model to your data.

Obtain Residuals: Calculate the residuals from the fitted model.

Generate Q-Q Plot: Plot the residuals against a theoretical normal distribution.

Assess Normality: Check if the residuals follow a straight line in the Q-Q plot.

Importance of Q-Q Plot in Linear Regression

Model Validation: Helps validate the assumption of normality of residuals. If the residuals are not normally distributed, the estimates of the coefficients, their standard errors, and the overall inference might be misleading.

Identifying Outliers: Outliers and heavy tails in the distribution can be easily identified if the points deviate significantly from the straight line.

Improving Model: If residuals are not normal, transformations of the dependent variable or the use of different modeling techniques might be necessary.

Interpretation

Straight Line: If the residuals fall approximately along the 45-degree reference line, it suggests that the residuals are normally distributed.

Deviations: Large deviations from the line, especially in the tails, indicate that the residuals are not normally distributed, suggesting potential issues with the model assumptions.

In linear regression analysis, the Q-Q plot is an essential diagnostic tool that helps make sure the residuals' normality assumption is satisfied. One may visually evaluate the residuals' degree of adherence to a normal distribution to make well-informed choices on the suitability of the linear regression model and, if required, take remedial action.